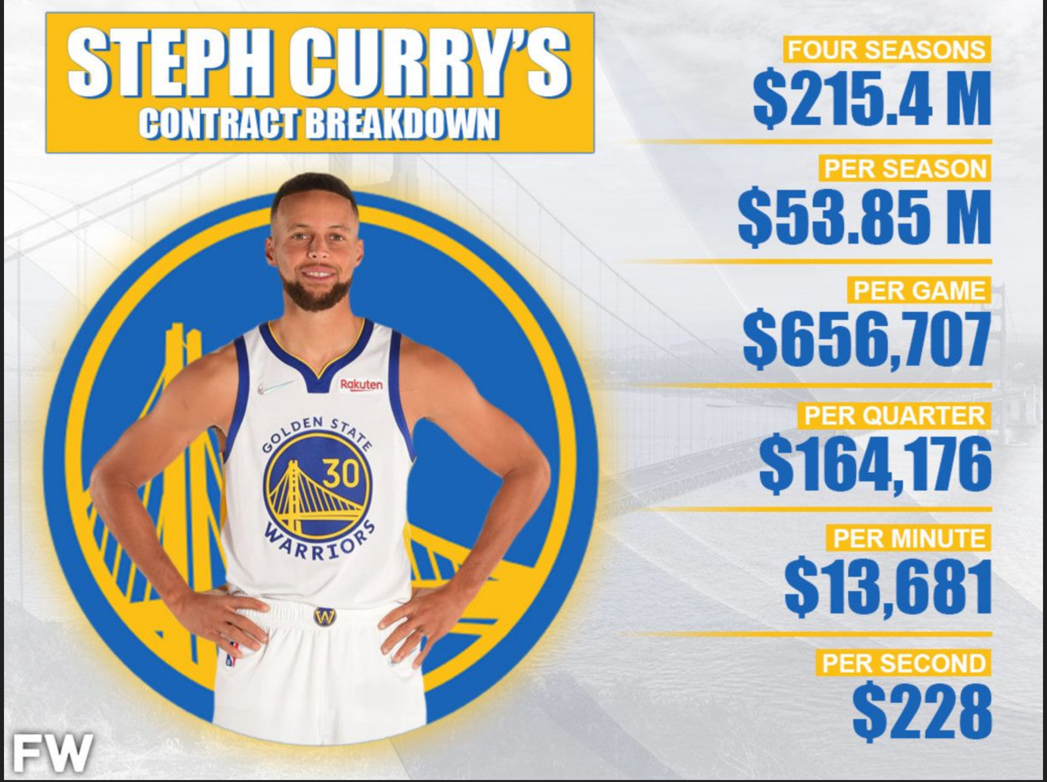
Capstone Presentation

NBA Salary Prediction and Assessment

Problem Statement



Data Collection

- 2020-2021 NBA Player Stats: Per Game
- 2020-2021 NBA Player Stats: Advanced
- NBA Contracts Summary
- Springboard Sports Database

Data Wrangling

- Merged data via player’s name

Data Wrangling

- Merged data via player’s name
- Dropped players with missing FG% variables.

Data Wrangling

- Merged data via player’s name
- Dropped players with missing FG% variables.
- Kept players with missing 3 point FG%

Data Wrangling

- Merged data via player’s name
- Dropped players with missing FG% variables.
- Kept players with missing 3 point FG%
- Kept players’ statistics on different team

Data Wrangling

- Merged data via player’s name
- Dropped players with missing FG% variables.
- Kept players with missing 3 point FG%
- Kept players’ statistics on different team
- 1221 rows x 58 columns to 454 rows * 51 columns

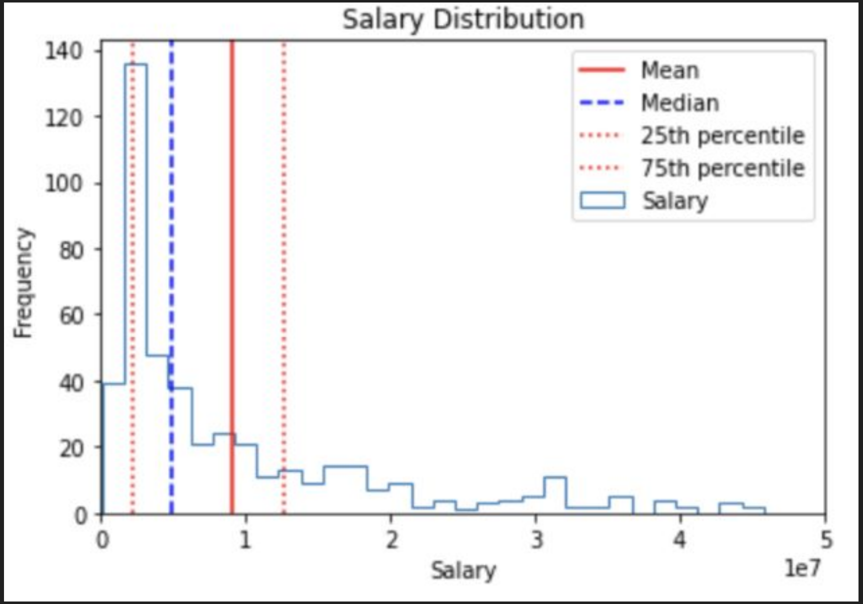
Exploratory Data Analysis

Certain insights:

- The average point scored by NBA players is around 10.65.
- The NBA is a young players driven league, with players under 25 occupying more than 50% of the league roster spots.
- While shooting percentages and 3 point percentages tend to follow a normal distribution, field goals made and attempts tend to be a shape that is skewed to the right.
- Salary is also skewed to the right, with certain bins popping at the very end

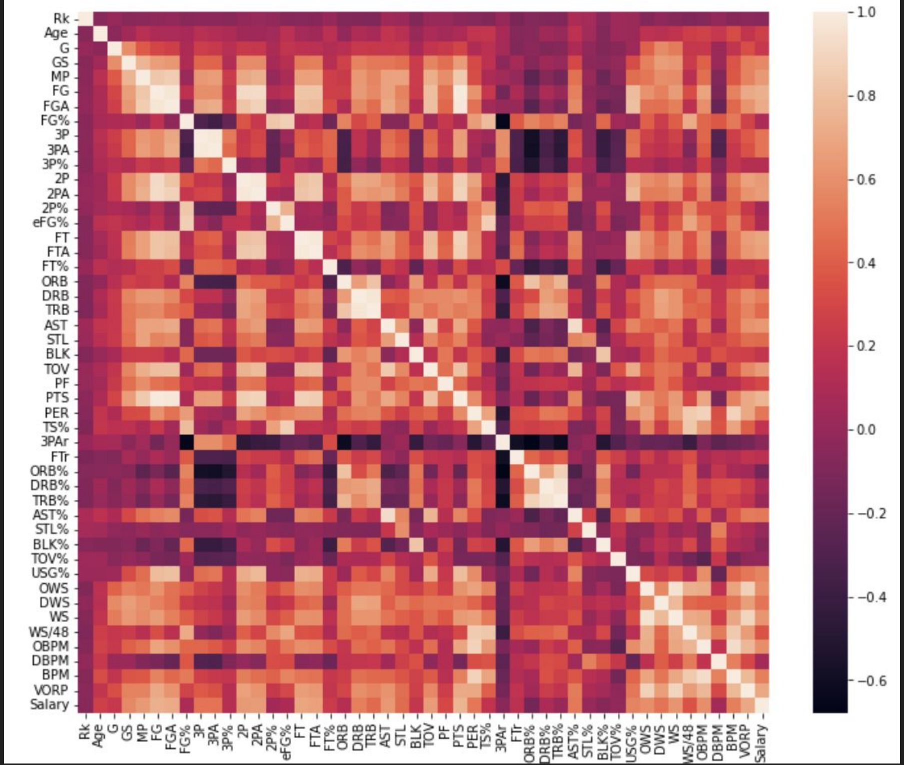
Exploratory Data Analysis

Salary Distribution



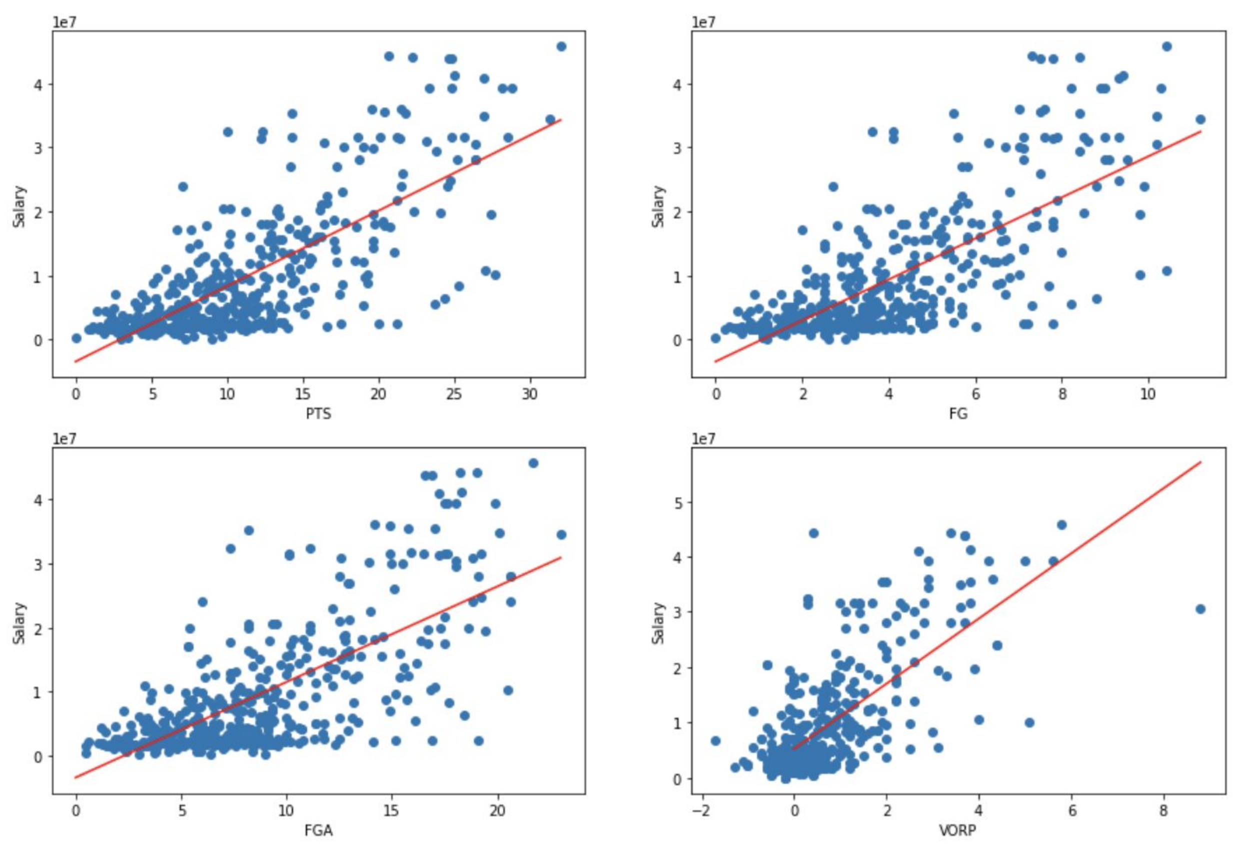
Exploratory Data Analysis

Heatmap for Salaries and other variables

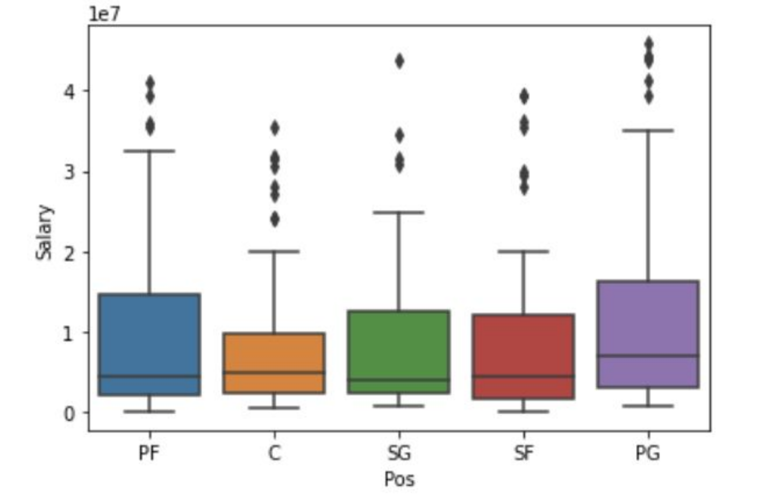


- Points
- FG makes
- FG attempts
- VORP (Value Over Replacement Players)

Exploratory Data Analysis

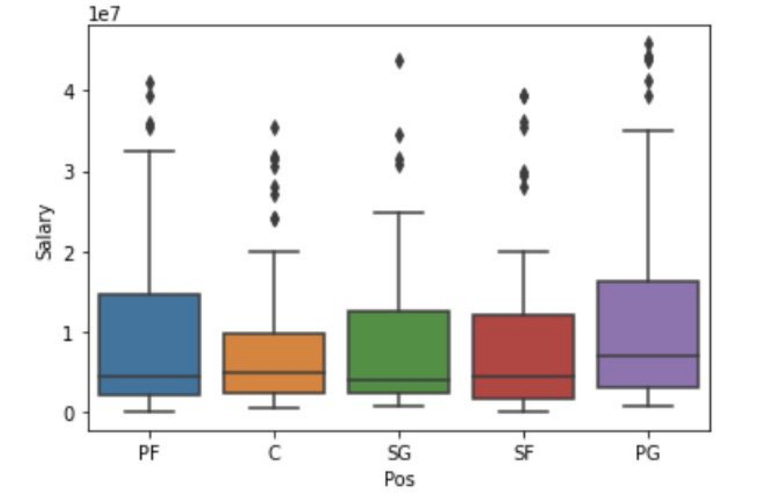


Categorical Data Treatment



- Keeping Positions as a Variable

Categorical Data Treatment



- Keeping Positions as a Variable

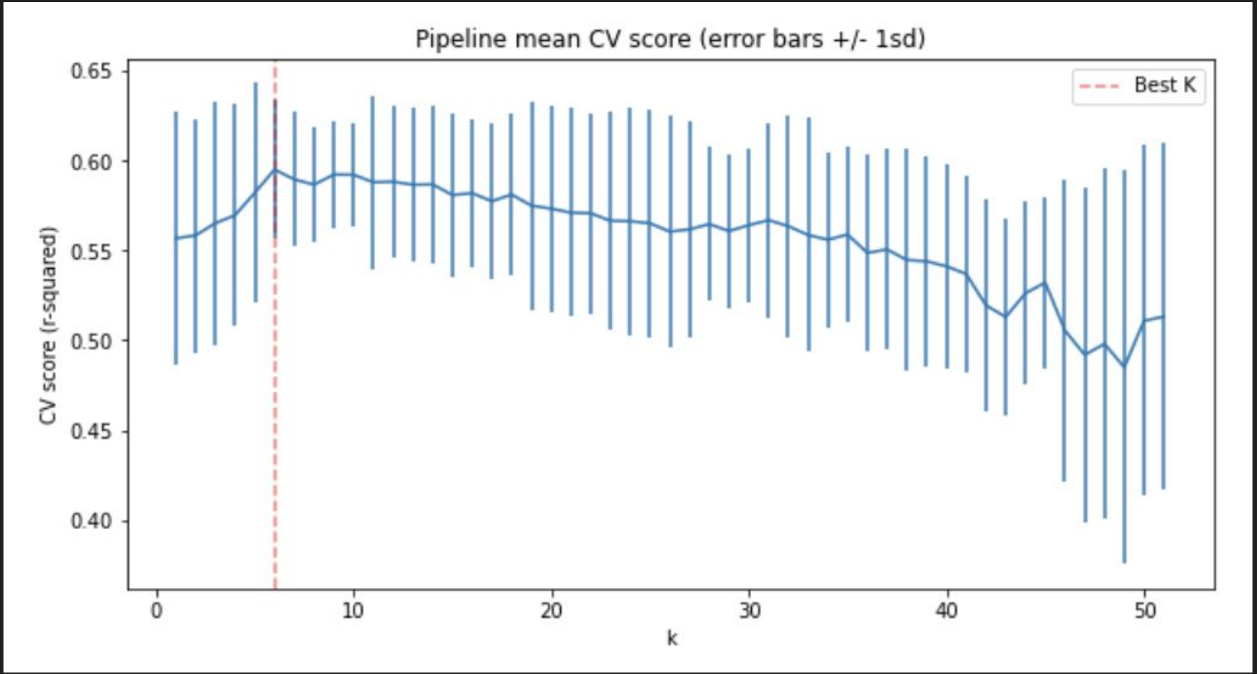
- **Null hypothesis: Being point guards do not have any impact on a player's salary. We reject the hypothesis with $p<0.05$.**

Modeling

- Linear Regression
- Random Forest Regression
- Gradient Boosting Regression

Linear Regression

- Select K-Best



Random Forest

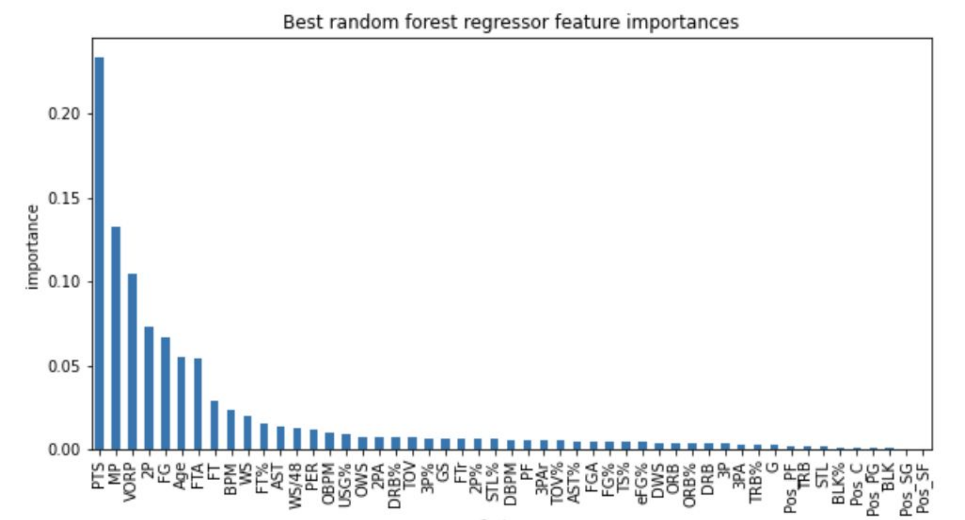
‘bootstrap': True

‘max_depth': 80,

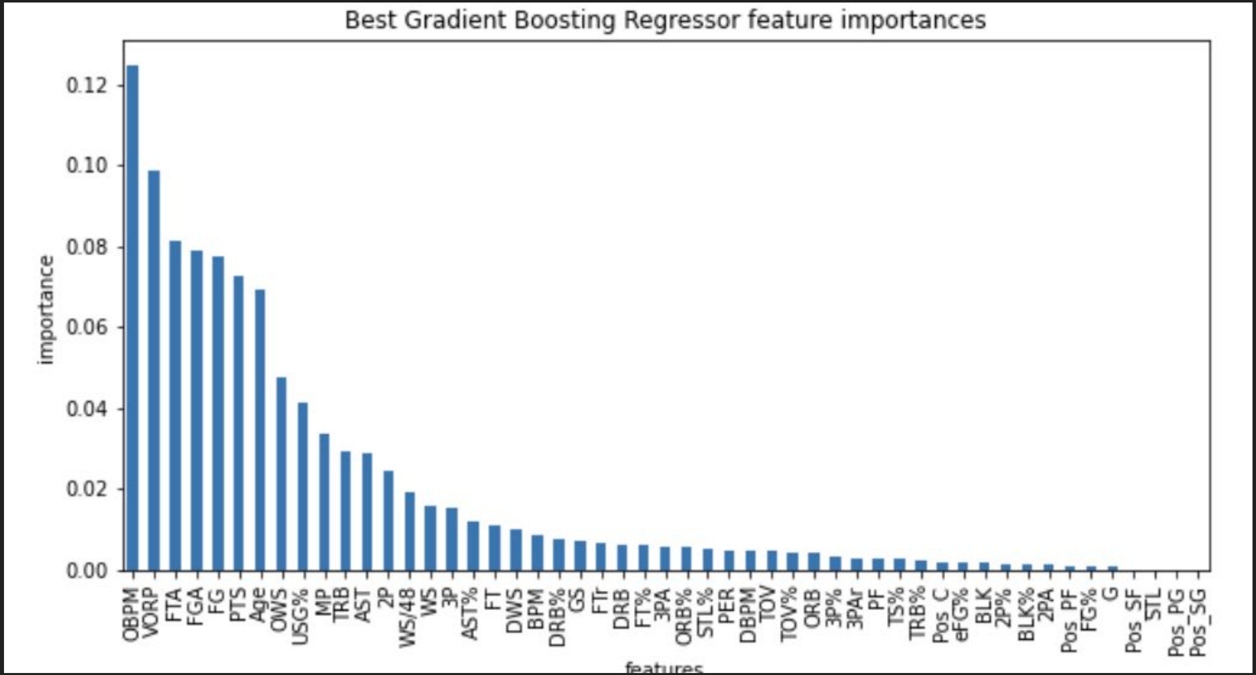
‘min_samples_leaf': 3,

‘n_estimators': 33,

StandardScaler()



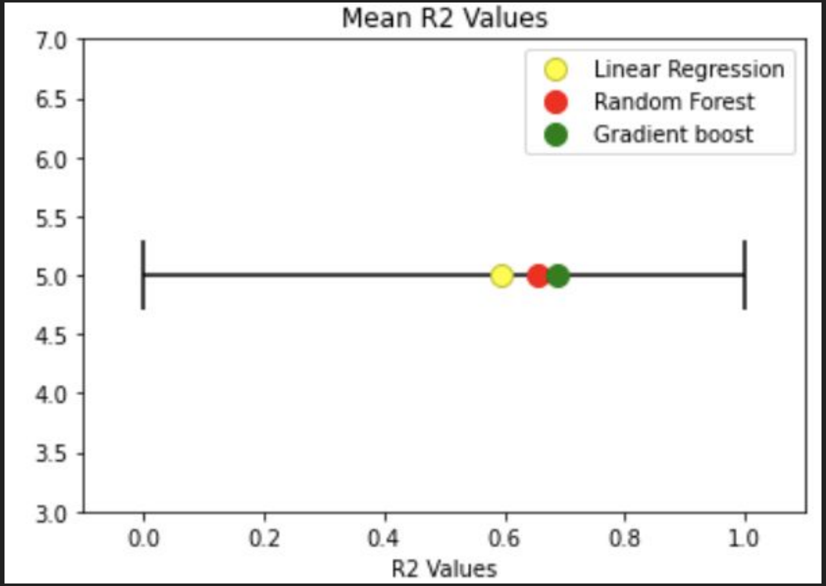
Gradient Boosting Regression



learning_rate
max_depth
max_features
n_estimators

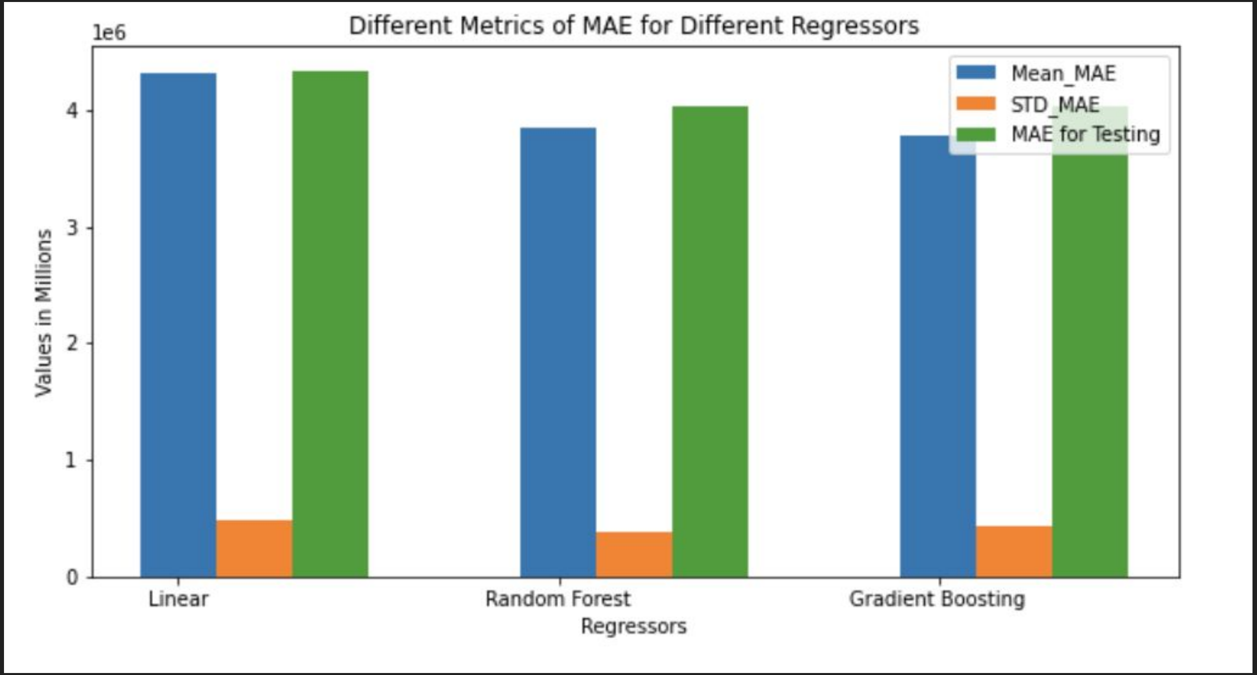
Final Assessment with 5 fold cross validation

- R Squared



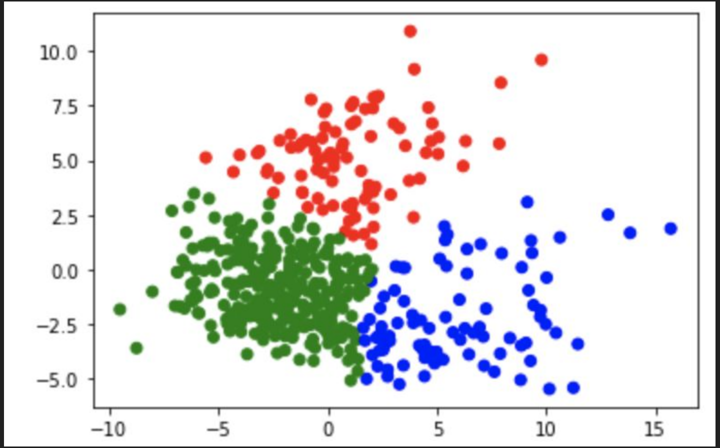
Final Assessment with 5 fold cross validation

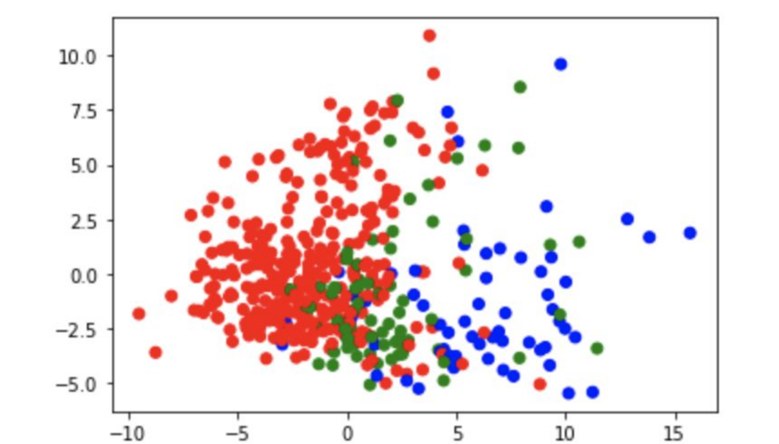
- Mean Absolute Error



PCA and K-Means

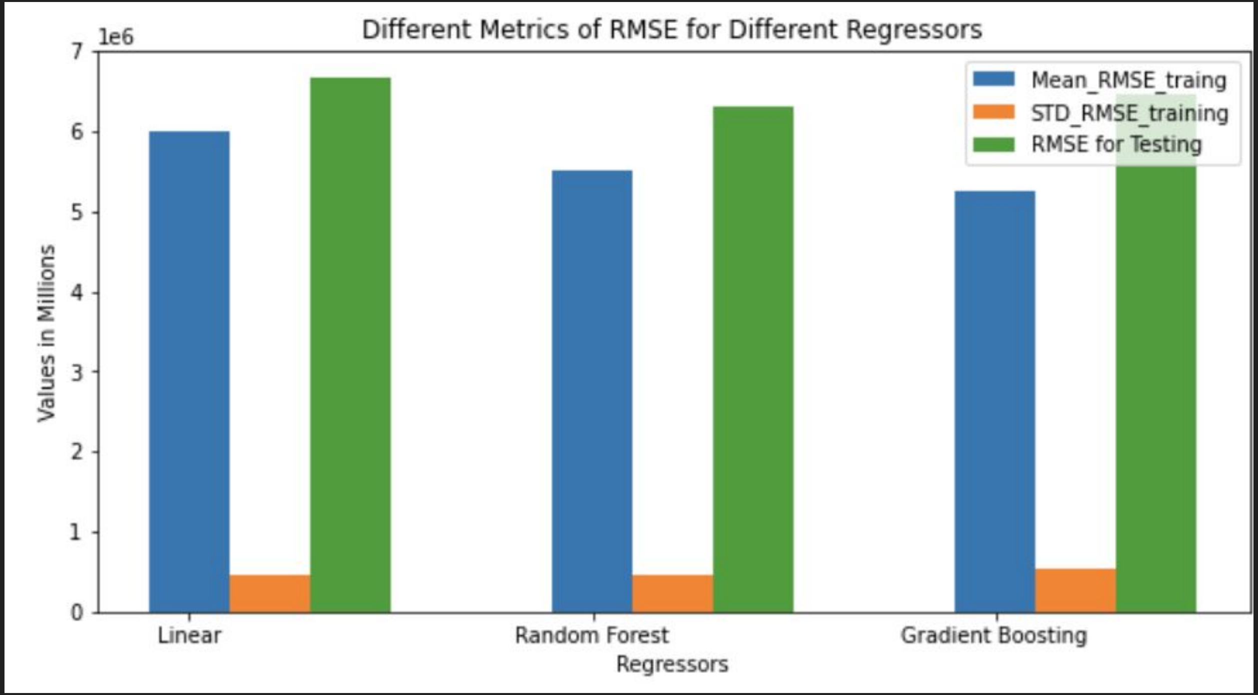
- 2 features and 3 clusters





Final Assessment with 5 fold cross validation

- Root Mean Squared Error



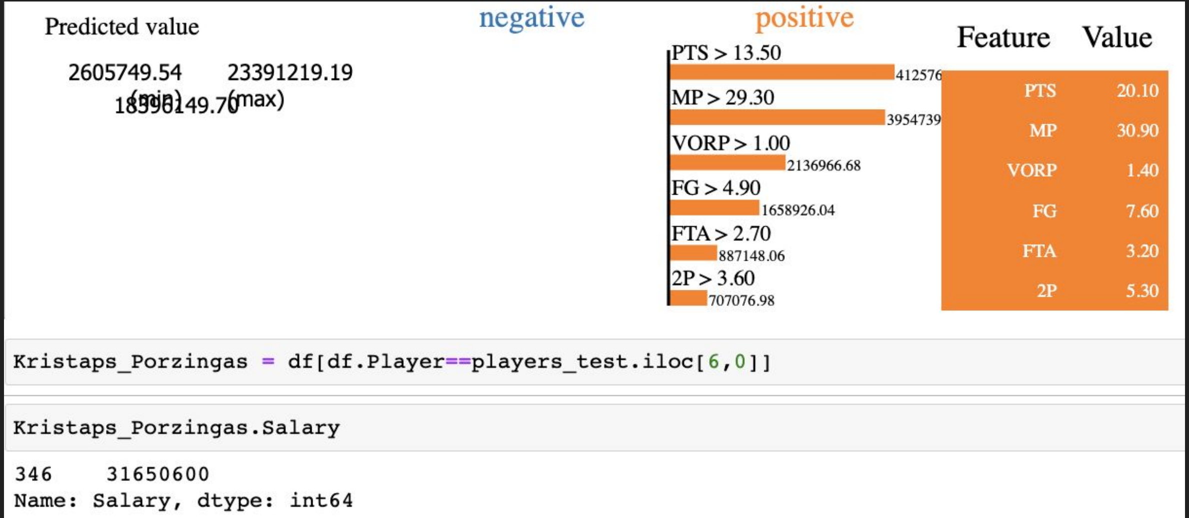
Final Model Selection and Reasoning

Random Forest Regression

- Performance on testing data
- Not overfitting

Lime Interpretation

Kristaps Porzingas



Feature	Value
PTS	20.10
MP	30.90
VORP	1.40
FG	7.60
FTA	3.20
2P	5.30

```
Kristaps_Porzingas = df[df.Player==players_test.iloc[6,0]]
```

```
Kristaps_Porzingas.Salary
```

```
346    31650600
Name: Salary, dtype: int64
```

How to Improve in the Future?

- Unsupervised Learning
- Classification Model